JOURNAL OF CHEMICAL INFORMATION AND MODELING

Subscriber access provided by Universitätsbibliothek Bern

Application Note

The Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning

Mahendra Awale, and Jean-Louis Reymond

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.8b00524 • Publication Date (Web): 17 Dec 2018

Just Accepted

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

The Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning

Mahendra Awale[a)] and Jean-Louis Reymond[a)]*

[a)] *Department of Chemistry and Biochemistry, National Center of Competence in Research NCCR TransCure, University of Berne, Freiestrasse 3, 3012 Berne Switzerland*

E-mail: jean-louis.reymond@dcb.unibe.ch

Abstract

Here we report PPB2 as a target prediction tool assigning targets to a query molecule based on ChEMBL data. PPB2 computes ligand similarities using molecular fingerprints encoding composition (MQN), molecular shape and pharmacophores (Xfp), and substructures (ECfp4), and features an unprecedented combination of nearest neighbor (NN) searches and Naïve Bayes (NB) machine learning, together with simple NN searches, NB and Deep Neural Network (DNN) machine learning models as further options. Although NN(ECfp4) gives the best results in terms of recall in a 10-fold cross-validation study, combining NN searches with NB machine learning provides superior precision statistics, as well as better results in a case study predicting off-targets of a recently reported TRPV6 calcium channel inhibitor, illustrating the value of this combined approach. PPB2 is available to assess possible off-targets of small molecule drug-like compounds by public access at ppb2.gdb.tools.

Introduction

In ligand-based virtual screening (LBVS) one uses structural similarities between molecules to infer possible similarities in their biological activities.[1] The approach is used broadly to accelerate and reduce the costs of the initial steps of drug discovery by guiding experimental screening to molecules with a higher probability of activity. LBVS also forms the basis for target prediction tools, which assign possible targets to a molecule based on its similarity to known, target annotated molecules such as those in the ChEMBL database.[2] Targets are assigned either directly based on nearest neighbor (NN) relationships, or indirectly by building a machine learning (ML) model,[3-22] with several tools available online,[23-34] with different levels of performance depending on the datasets, models and hyperparameters used.[35-39] Such target prediction tools are essential to help assess the polypharmacology of any hit compound or drug molecule.

We recently reported PPB (Polypharmacology Browser), a public web-portal for target prediction which used simple NN searches with multiple fingerprints simultaneously on ChEMBL data to predict off-targets.[34] While showing good performance in benchmarking studies, PPB failed to correctly predict hERG as an important off-target of CIS22a, a TRPV6 inhibitor recently reported by our group,[40] a limitation shared by other online target prediction tools. This motivated us to rethink our approach, which led us to PPB2 (Polypharmacology Browser version 2) presented herein, featuring a redesigned approach to target prediction. While our previous tool PPB used all available ChEMBL data including proteins, cell lines and organisms, we focused on protein targets only for the new tool PPB2 to have fewer but better defined cases. Furthermore, in contrast to PPB which exploited six different fingerprints and four fusion fingerprints using nearest neighbor (NN) searches, for PPB2 we selected only three fingerprints from PPB encoding different levels of detail, including the best performing ECfp4 (extended connectivity fingerprint up to four bonds),[41] but used them in a combination of NN searches with Naïve Bayes (NB) machine learning. To the best of our knowledge this approach is unprecedented for target prediction, although a related method combination has been reported previously for QSAR studies.[42] For comparison PPB2 also included

NN searches similar to those of PPB, as well as NB and Deep Neural Network (DNN) machine learning models.

Our results show that while NN searches based on ECfp4 is the best method in PPB2 in terms of recall in a 10-fold cross-validation study, the combination of NN searches with NB machine learning perform best in terms of precision statistics. The combination method also stands out in a case study predicting off-targets of a recently reported TRPV6 calcium channel inhibitor,[40] where they correctly predict hERG as an important off-target missed by NN searches as well as by our previous tool PPB and by many other online target prediction tools. PPB2 is freely available for use at ppb2.gdb.tools.

Results and Discussion

To build PPB2 we collected a bioactivity dataset of all compounds having at least $IC_{50} < 10$ μM on a single protein target in ChEMBL22,[43] considering only high confidence data points as annotated in ChEMBL and only targets for which at least 10 compounds were documented. This provided 344,163 single compounds associated with 1,720 single protein targets belonging to 8 different target families, representing 555,346 target-compound associations.

To encode molecular structures we selected three fingerprints from PPB perceiving different levels of details, namely: 1) MQN (Molecular Quantum Number), a 42-bit scalar fingerprint representing molecular composition by atoms, bonds, polar groups and topological features[44, 45] particularly useful to search and visualize large databases;[46-48] 2) Xfp (atom category extended atom-pair fingerprint), an 55-bit scalar fingerprint perceiving molecular shape and pharmacophores and well suited for scaffold-hopping virtual screening;[49] and 3) ECfp4 (extended connectivity fingerprint up to four bonds), a 1024-bit binary substructure fingerprint encoding detailed information about molecular structure,[41] and which performed best in PPB. Similarities were computed using the city-block distance for MQN and Xfp and the Tanimoto coefficient for ECfp4.[50]

In terms of search methods, we first implemented NN searches with each of these three fingerprints. Note that in contrast to PPB where the final target list was re-ordered according to p-values obtained from NN similarities by calculating the probability of a similarity value to occur at random, we used the direct order of targets given by NN similarities. This simplification avoided any bias resulting from the choice of reference molecules used to create the random distribution (which for PPB were molecules from ZINC).

Secondly, we implemented our key idea for PPB2, which was to build a specific ML model for each query molecule using NN compounds of the query. This approach would avoid making predictions based on non-realistic associations between compounds that would be structurally too distant from each other, thus capitalizing on both the "no-nonsense" advantage of NN searches and the deeper exploitation of multiple datapoints possible with ML. The best combination involved retrieving the 2,000 NN of a query molecule using MQN, Xfp, or ECfp4, followed by building a Laplacian modified Multinomial NB model from these 2,000 NN to provide the actual target prediction. The number of 2,000 NN was large enough to include all high-similarity compounds and small enough to enable a fast NB model building for each query molecule on the fly. Finally, we also implemented a direct NB model and a deep neural network (DNN) model trained with the entire dataset using ECfp4, the best performing fingerprint from our previous tool PPB. In total this provided eight different search methods for PPB2 (Figure 1a).

In the PPB2 web-portal the user inputs a molecule as either SMILES or structural drawing and selects a prediction method (Figure 1b). Prediction results are shown as a list of the most probable targets (Figure 1c). For each target the ChEMBL molecules on which the target prediction is based can be inspected visually by opening the "Show NN" window. Molecular structures are drawn from SMILES within the browser using SmilesDrawer (Figure 1d).[51] A tutorial and a FAQ (frequently asked question) tabs are available with detailed explanation on the methods and tool.

The performance of the eight different target prediction methods in PPB2 was analyzed by calculating recall and precision statistics of a 10-fold cross-validation study (see methods for details). The performance in terms of overall recall (percentage of known compound-target pairs in the test set that are predicted by PPB2) over the eight methods tested increased from 47-61 % when considering only the single top predicted target, to 85-96 % when considering the ten top predicted targets (Figure 2a, top left panel). At the same time the overall precision (percentage of compound-target pairs predicted by PPB2 which correspond to known interactions) decreased from 69-91 % when considering only the single top predicted target, to 13-14 % when considering the ten top predicted targets (Figure 2a, bottom left panel). Recall was lower but precision higher with a similar trend when calculated on a per target basis, which compensates for the fact that some targets are overrepresented (Figure 2a, top right and bottom right panels). Similar results were obtained when considering only compounds associated with a single protein target (Figure 2b). Performance was most strongly influenced by the ECfp4 Tanimoto similarity between the query molecule and its ECfp4 NN in the training set, with a drastic reduction of both recall and precision when the value dropped below 0.5 (Figure 2c/d). Performance was also influenced by the target class, with membrane receptors performing best and kinases performing worst, an effect also observed with other target prediction tools (Figure 2e).[13]

Across the different comparisons above, the simple NN(ECfp4) method, used already in PPB, performed best in terms of recall statistics, which has also been noted by other authors in a related target prediction study.[17] In terms of precision statistics by contrast, there was a clear advantage for the combined NN+NB methods, in particular when computed on a per target basis, which is an important parameter as it gives an indication of the expected rate of success for target assignment (lower right panels in Figure 2a-d and Figure 2e). On the other hand, the simple NN(MQN) and the NB(ECfp4) methods gave relatively poorer results across all comparisons, showing that encoding compounds by composition only (MQN) or using a NB model alone (even

with ECfp4) does not exploit the target-compound pair information as well as the other, more complex methods in PPB2 for target prediction.

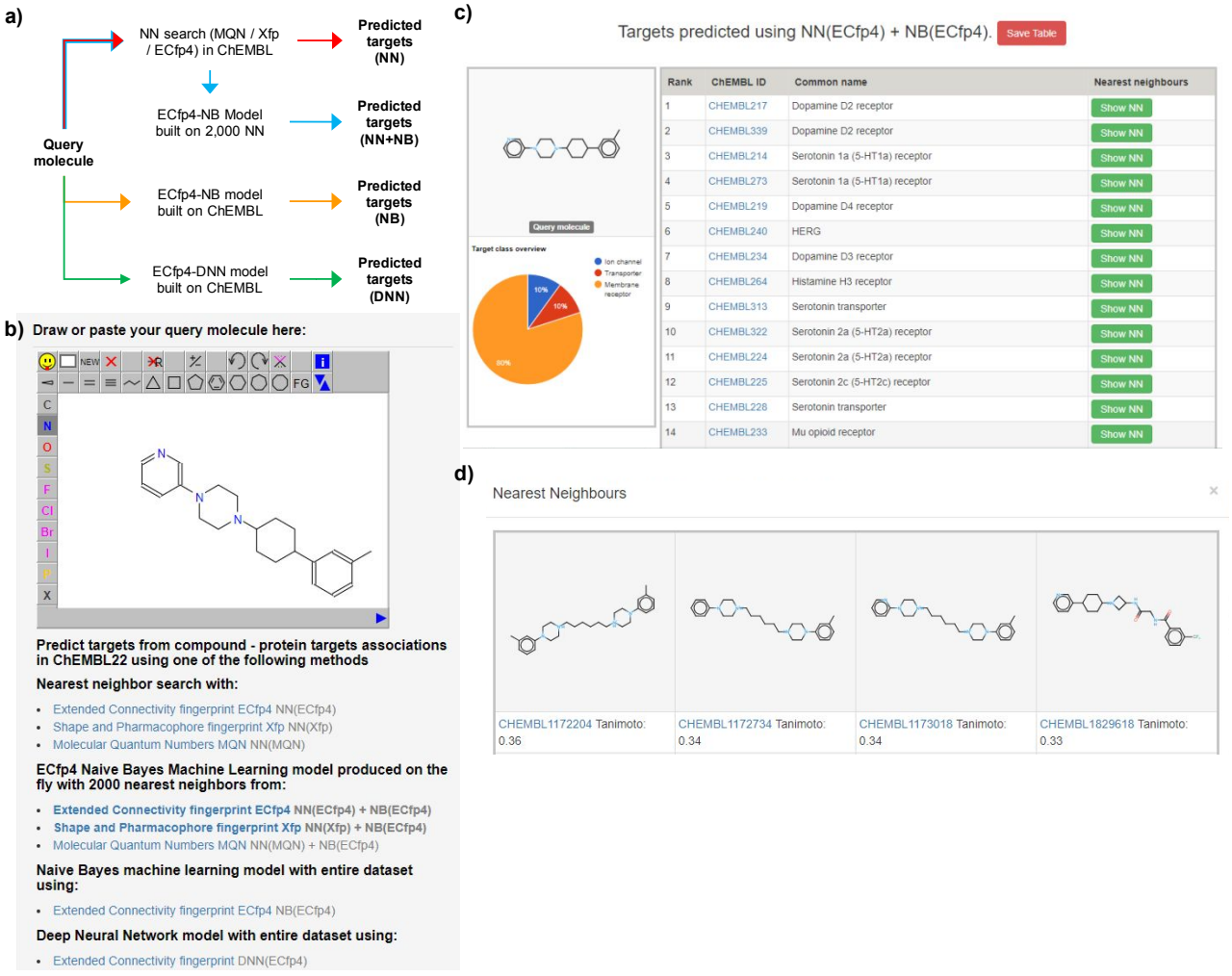

Figure 1. The polypharmacology browser PPB2. **(a)** Workflow of the four different target prediction methods. Red arrows: retrieve top 2,000 NN using MQN (city block distance) or Xfp (city block distance) or ECfp4 (tanimoto coefficient) and rank the targets as per the similarity score of the most similar nearest neighbor associated with each target. Blue arrows: retrieve top 2,000 NN using MQN or Xfp or ECfp4, then build an ECFP4-Naïve Bayes (NB) machine learning model based on these 2,000 nearest neighbors and perform the target prediction. Orange arrows: perform target prediction using an ECfp4-Naïve Bayes machine learning model built on ChEMBL. Green arrows: perform target prediction using an ECfp4-Deep Neural Network (DNN) machine learning model built on ChEMBL. **(b)** Query molecule entry window in PPB2 with example molecule CIS22a. **(c)** PPB2 target prediction results window showing rank, ChEMBL ID and common name for predicted targets, exemplified with compound CIS22a and the NN(ECfp4) + NB(ECfp4) method. **(d)** Display of the first four NNs associated with hERG. This panel opens through the "Show NN" green button in the target list (c).

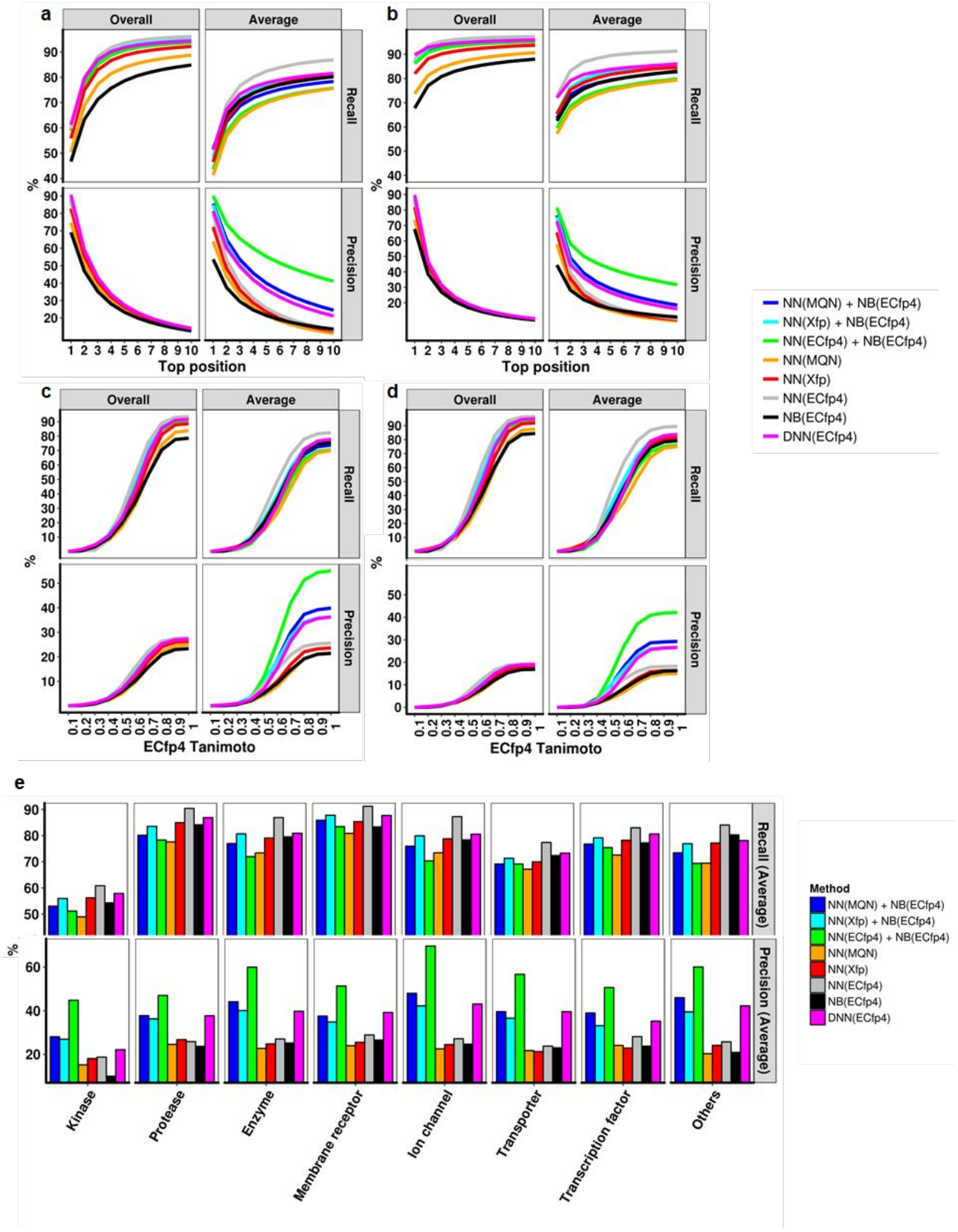

Figure 2. Performance of PPB2 in a 10-fold cross-validation (see method for details). **(a)** Overall (all ligand-target pairs) and average (per target) recall and precision performance as function of the number or targets considered. **(b)** Same as (a) considering only compounds annotated with a single target (241,083 compounds, ~70% of database). **(c)** Overall and average recall and precision considering top 5 predicted targets as function of ECfp4 tanimoto coefficient similarity of the compound in the test set to its NN in the training set. **(d)** Same as (c) for single target compounds only. **(e)** Average performance (per target) estimated per target class considering the top 5 predicted targets.

As an application example we analyzed the case of compound CIS22a, a recently reported TRPV6 calcium channel inhibitor for which we had measured 24 possible off-targets from the "safety screen" panel of Cerep Pvt. Ltd.[40] We compared the prediction of PPB2 with those of SwissTargetPrediction,[29] SEA,[24] Modlab Spider,[28] SuperPred,[30] HitPick,[27] TargetHunter,[25] Chemmapper,[26] TarPred,[32] PASS,[31] as well as our previously reported PPB, considering for each tool the top 20 predicted targets (Figure 3, the structure of CIS22a is shown in Figure 1).[34]

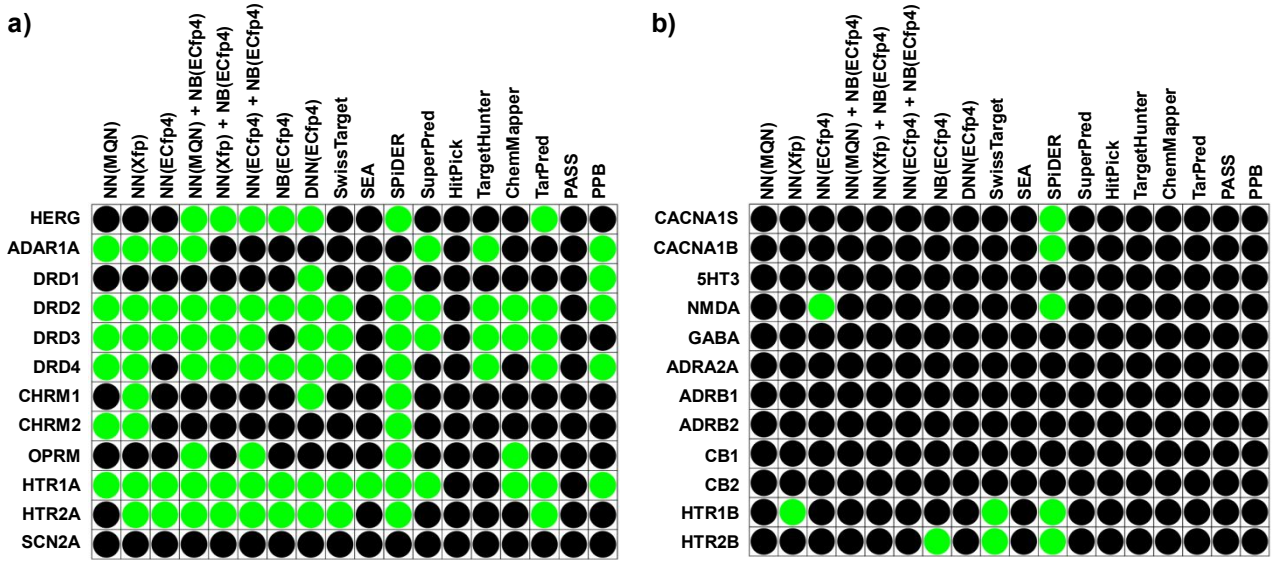

Figure 3. Target prediction results for the eight prediction methods in PPB2 and other prediction tools available online for the TRPV6 inhibitor CIS22a. The structure of CIS22a is shown in Figure 1. **a)** 12 off-targets against which CIS22a was found experimentally to be active. **b)** 12 off-targets against which CIS22a was found experimentally to be inactive (active is defined as > 50 % inhibition at 10 μM, and inactive is defined as ≤ 50 % inhibition at 10 μM, see supporting information of ref. no 39). For each target prediction tool targets are marked as green if predicted by the tool, or black if not predicted by the tool. For each option in PPB2 and each online tool predictions were made by considering the 20 top predicted targets. Target full names: Voltage gated potassium channel subfamily H member 2 (HERG), Adrenergic α1A (ADRA1A) and α2A (ADRA2A) receptor, Dopamine receptor subtypes D1-4 (DRD1-4), Cholinergic muscarinic receptor 1 (CHRM1) and 2 (CHRM2), μ opioid receptor (OPRM), 5-Hydroytryptamine receptor 1A (HTR1A), 1B (HTR1B), 2A (HTR2A) and 2B (HTR2B), Voltage gated Na+ channel (SCN2A), Voltage dependent L- (CACNA1S) and N-type (CACNA1B) Ca2+ channel, 5-Hydroytryptamine receptor 3 (5-HT3), N-methyl-D-aspartate receptor (NMDA), Gamma aminobutyric acid receptor (GABA), Adrenergic β1 (ADRB1) and β2 (ADRB2) receptor, Cannabinoid 1 (CB1) and 2 (CB2) receptor.

PPB2 predicted 6-8 of the 12 off-targets against which CIS22a is active, and only 1-2 of 12 possible off-targets against which CIS22a is not active. Most interestingly, the three combined NN+NB(ECfp4) methods as well as the NB(ECfp4) and DNN(ECfp4) correctly predicted hERG, an important off-target of CIS22a which was missed by the simple NN searches, illustrating the value of ML approaches. Note that the combined NN+NB methods as well as the DNN did not

predict any of the inactive off-targets, while the NB method incorrectly predicted one of the inactive off-targets.

Among the other search tools, only SPIDER and TarPred predicted hERG for CIS22a. TarPred also correctly predicted five other off-targets and did not predict any of the inactive off-targets. SPIDER predicted 11 out of the 12 off-targets of CIS22a including hERG, but also predicted 5 out of the 12 inactive off-targets, suggesting that this tool favors recall over precision. The other tools tested failed to predict hERG but did not predict any of the inactive off-targets, indicating that they favor precision over recall.

Conclusion

Here we redesigned our online target prediction tool PPB to a new version PPB2 by implementing three important changes: 1) focusing the dataset used for prediction protein targets by excluding cell lines and organisms to have fewer but better defined cases; 2) rather than the 10 fingerprints used in PPB with NN searches only, selecting only three fingerprints encoding different level of detail; 3) complementing NN searches with a combination of NN searches and Naive Bayes (NB) machine learning, providing NB and deep neural network (DNN) with ECfp4 as additional options, resulting in eight search options.

Remarkably, all eight methods in PPB2 performed well in our 10-fold cross-validation study, giving performance values comparable or better than values reported in the literature for other target prediction tools. While NN(ECfp4) performed best in terms of recall statistics, the combined NN+NB methods, which is the key innovation in PPB2, showed superior performance in terms of precision. The combined methods also returned better results in a case study with the TRPV6 calcium channel inhibitor CIS22a, in particular regarding prediction of hERG as off-target. PPB2 is freely available at ppb2.gdb.tools and can be used to assess possible off-targets of small molecule drug-like compounds.

Methods

Compound-target interactions database

All compound-target interactions were extracted from the publicly available MySQL version of the ChEMBL database (release 22). Criteria for selecting interactions from the database were targets labeled as "single protein" with the source organism being either human or rat and compounds associated with the target were to have IC_{50}, EC_{50}, K_i or K_D value of ≤ 10 μM, confidence score > 5 and heavy atom count ≤ 50. In a next step, compounds were processed in non-isomeric SMILES format, where counter ions were removed, valence errors corrected, and the compounds ionized at pH 7.4. Duplicate molecules were then removed based on unique SMILES comparison in the context of each target protein, and targets with less than 10 compounds were discarded. Filtering ChEMBL22 using these criteria resulted in a subset containing 1,720 target proteins, 344,163 unique compounds and 555,346 compound-target interactions. This subset was stored in a plain text file (henceforth referred to as the database) grouped by SMILES resulting in 344,163 lines, with each line containing a compound encoded as a SMILES string followed by the associated target names.

Fingerprints

For each compound in the database we calculated the MQN (42D), Xfp (55D) and ECfp4 (1024D) fingerprints. All fingerprint calculations were performed using an in-house java program utilizing various plugins from the ChemAxon JChem library such as MajorMicrospeciesPlugin to adjust the ionization state of molecules, HBDAplugin to determine hydrogen bond donor and acceptor atoms and TopologyAnalyserPlugin to determine the shortest topological path between atom-pairs.

NN searches

For a given query molecule, NN searches are performed in the database using MQN, Xfp, or ECfp4 fingerprint. For MQN and Xfp, the similarity between query (A) and a compound in the database

(B) is calculated using city block distance (CBD) as follows, where A = fingerprint of the query molecule, B = fingerprint of a compound in the database:

$$CBD(A,B) = \sum_{i=1}^{n} |A_i - B_i|$$

For the ECfp4 fingerprint the similarity between compounds is calculated using the Tanimoto coefficient (T) as follows, where A = ECfp4 fingerprint of the query molecule, B = ECfp4 fingerprint of a compound in the database, N_C = number of ON bits common between molecule A and B, N_A *and* N_B = number of ON bits in the query and a compound in the database, respectively:

$$T(A, B) = \frac{N_C}{N_A + N_B - N_C}$$

After the similarity calculation compounds in the database are sorted with respect to the query molecule (high to low similarity: increasing CBDs or decreasing Tanimoto coefficient) and the top 2,000 compounds are extracted. Targets associated with these top 2000 compounds are then extracted and sorted as per similarity score of the closest nearest neighbor associated with a target.

NB Model

The Naive Bayes model was created using ECfp4 fingerprints of all the compounds in the database. A Multinomial Naïve Bayes Classifier (NB) model with Laplacian smoothing was trained using the Python-based scikit-learn (version 0.19) machine learning library. Fingerprints of molecules were used as input feature vectors, while target names were used as class labels. The NB classifier is probabilistic in nature, which means that for any given query compound the NB classifier calculates the probability for each target present in the database and considers the targets with the highest probabilities (top N targets) as the predicted targets of a query molecule. This probability calculation is fundamentally based on the Bayesian theorem of conditional probability, wherein the probability of an event A happening considering an evidence B ($P(A \mid B)$) can be calculated based on prior probability of an evidence B in the sample of an event A ($P(B \mid A)$) and prior probability of

an event A ($P(A)$) in the training dataset. For target prediction (A = target, B = fingerprint of a molecule) the probability $P(A \mid B)$ of a compound to be active on target A is calculated from ($P(B \mid A)$) = prior probabilities calculated based on fingerprints of known bioactive compounds active against target A in the database, $P(A)$ = the relative frequency of target A considering all the other targets in the database, and $P(B)$ = the probability of evidences across the entire database, as follows:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

The NB classifier uses the naïve assumption that all the features (evidences) are independent from each other and contribute independently to probability calculation. To calculate $P(A \mid B)$ given multiple evidences ($B = B_{i=1}...B_n$) one needs to estimate the product of the probability of each evidence given an event A ($P(B \mid A) = P_{B1A} \cdot P_{B2A}.. \cdot P_{nA}$) and multiplied by the prior probability of an event A in the database. The probability for each of the evidences given event A is True (P_{BiA}) can be calculated as follows, where N_{iA} = count of i^{th} evidence in a sample of event A in the database, N_A = total count of all evidences in a sample of an event A in the database. n = number of evidences and α = smoothing factor:

$$P_{BiA} = \frac{N_{iA} + \alpha}{N_A + \alpha n}$$

When $\alpha = 1$, the calculation is called Laplace smoothing. Parameter α is useful to avoid zero probability resulting from the absence of the evidences in the sample. Further technical details about the NB classifier can be found in the Scikit-learn documentation and corresponding GitHub repository (https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/naive_bayes.py)

NN searches combined with NB Model

For any given query molecule, the top 2000 NN are extracted from the database using MQN, Xfp, or ECfp4. A Naïve Bayes machine learning model is then built on the fly using ECfp4 fingerprints

of these 2000 NN as feature vectors and target names as class labels. This NB model, which is specific to the query molecule, is then used to predict the targets of the query compound.

Deep Neural Network Model

The deep neural network model was created using ECfp4 fingerprints of all the compounds in the database. The DNN model was built using Keras (version 2.0.9), a python based deep learning library with Tensorflow-gpu backend. The DNN model presented herein contains a 1,024D input layer corresponding to 1,024D ECfp4 fingerprints of molecules, two hidden layers containing 1,000 and 500 neurons respectively, and a 1,720D binary vector as output layer (1: compound active against target, 0: compound not active against target) corresponding to 1,720 targets in the database. Rectified linear units (relu) were used as activation functions for the hidden layers with a drop-out rate of 20% to avoid model-overfitting, while a sigmoid activation function was used for the output layer. The model was trained using "adam" as optimizer and binary crossentropy as loss function. The numbers of epoch iterations and the batch size were set to 200 and 500 respectively. Initially, series of models were built and evaluated in 10-fold cross validation with progressive optimization of number of neurons and number of hidden layers in network.

PPB2 website

PPB2 front-end is implemented using Html, JavaScript and Bootstrap front-end web framework, while the PPB2 back-end is implemented using Flask Python web framework. PPB2 is publicly accessible at http://ppb2.gdb.tools/. The target prediction models/methods implemented in PPB2 website are based on the entire database**.**

Recall and Precision in cross-validation study

The target prediction performance of the different methods (see Figure 1) was measured using recall and precision parameters in a 10-fold cross validation study. For this study the records in the database (lines in the text file), were randomized and subdivided into 10 distinct sets of equal size.

Each of these 10 sets was then used as a test dataset with the remaining 9 sets being combined into the training dataset, replacing the database in the descriptions above. We carried out target prediction only for compounds associated with up to 10 targets (this corresponds to 342,706 compounds out of 344,164 total compounds and 508,153 interactions out of 555,346 total interactions). For each compound in the test set we considered the top *n* predicted targets and evaluated whether the annotated targets for the compound could be recalled by the prediction. For each target *x* we calculated a) true positives (TP): the number of times target *x* was correctly predicted; b) false negatives (FN): the number of interactions of target *x* which were not predicted; c) false positives (FP): the number of times target *x* was wrongly predicted (found among top *n* predicted targets but not annotated for the compound).

We calculated overall recall, average recall, overall precision and average precision as follows: *Overall recall:* The number of true positives divided by the number of true positives plus the number of false negatives across all the targets (*i*) and all the 10 cross-validation runs (*j*) as shown in equation 1. *Average recall:* The average recall for each target across 10 cross-validation runs (equation 2). Similarly, the overall precision (equation 3) and average precision (equation 4) were calculated. Furthermore, targets were classified as per target class and overall recall, average recall, overall precision and average precision were calculated class-wise.

$$Overall\ recall = \frac{\sum_{i=1}^{n}\sum_{j=1}^{10} TP_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{10} TP_{i,j} + \sum_{i=1}^{n}\sum_{j=1}^{10} FN_{i,j}} \cdot 100 \quad (1)$$

$$Average\ recall = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{10} TP_{i,j}}{\sum_{j=1}^{10} TP_{i,j} + \sum_{j=1}^{10} FN_{i,j}}\right) \cdot 100 \quad (2)$$

$$Overall\ precision = \frac{\sum_{i=1}^{n}\sum_{j=1}^{10} TP_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{10} TP_{i,j} + \sum_{i=1}^{n}\sum_{j=1}^{10} FP_{i,j}} \cdot 100 \quad (3)$$

$$Average\ precision = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{10} TP_{i,j}}{\sum_{j=1}^{10} TP_{i,j} + \sum_{j=1}^{10} FP_{i,j}}\right) \cdot 100 \quad (4)$$

Where $TP = true\ positives$, $FP = false\ positives$, $FN = false\ negatives$, $i = target$,

$$j = cross\ validation\ run$$

Authors' contributions. MA designed, realized and implemented PPB2 and wrote the paper, JLR designed supervised the study and wrote the paper.

Notes. The authors declare no competing financial interest.

Acknowledgments

This work was supported financially by the University of Berne, the Swiss National Science Foundation and the NCCR TransCure. We thank ChemAxon Pvt. Ltd. for providing free academic and web licenses for their products.

References

1. Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012,** *52*, 867-881.

2. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. Chembl: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012,** *40*, D1100-D1107.

3. Jenkins, J. L.; Bender, A.; Davies, J. W. In Silico Target Fishing: Predicting Biological Targets from Chemical Structure. *Drug Discov. Today Technol.* **2006,** *3*, 413-421.

4. Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006,** *46*, 1124-1133.

5. Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: "Target Fishing" Using 2d and 3d Molecular Descriptors. *J. Med. Chem.* **2006,** *49*, 6802-6810.

6. Cleves, A. E.; Jain, A. N. Robust Ligand-Based Modeling of the Biological Targets of Known Drugs. *J. Med. Chem.* **2006,** *49*, 2921-2938.

7. Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* **2007,** *2*, 861-873.

8. Dariusz, P.; Marcin von, G.; Stephane, A. H. S.; Leszek, R.; Lucjan, S. W.; Krzysztof, G.; Uwe, K. Target Specific Compound Identification Using a Support Vector Machine. *Comb. Chem. High Throughput Screen.* **2007,** *10*, 189-196.

9. Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B. O. Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* **2008,** *48*, 2313-2325.

10. Koutsoukas, A.; Simms, B.; Kirchmair, J.; Bond, P. J.; Whitmore, A. V.; Zimmer, S.; Young, M. P.; Jenkins, J. L.; Glick, M.; Glen, R. C.; Bender, A. From in Silico Target Prediction to Multi-Target Drug Design: Current Databases, Methods and Applications. *J. Proteom.* **2011,** *74*, 2554-2574.

11. AbdulHameed, M. D. M.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring Polypharmacology Using a Rocs-Based Target Fishing Approach. *J. Chem. Inf. Model.* **2012,** *52*, 492-505.

12. Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Ritchie, D. W. Detecting Drug Promiscuity Using Gaussian Ensemble Screening. *J. Chem. Inf. Model.* **2012,** *52*, 1948-1961.

13. Koutsoukas, A.; Lowe, R.; KalantarMotamedi, Y.; Mussa, H. Y.; Klaffke, W.; Mitchell, J. B. O.; Glen, R. C.; Bender, A. In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* **2013,** *53*, 1957-1966.

14. Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. Tools for in Silico Target Fishing. *Methods* **2015,** *71*, 98-103.

15. Afzal, A. M.; Mussa, H. Y.; Turner, R. E.; Bender, A.; Glen, R. C. A Multi-Label Approach to Target Prediction Taking Ligand Promiscuity into Account. *J. Cheminform.* **2015,** *7*, 1-14.

16. Lusci, A.; Fooshee, D.; Browning, M.; Swamidass, J.; Baldi, P. Accurate and Efficient Target Prediction Using a Potency-Sensitive Influence-Relevance Voter. *J. Cheminform.* **2015,** *7*, 1-13.

17. Czodrowski, P.; Bolick, W.-G. Ocean: Optimized Cross Reactivity Estimation. *J. Chem. Inf. Model.* **2016,** *56*, 2013-2023.

18. Lavecchia, A.; Cerchia, C. In Silico Methods to Address Polypharmacology: Current Status, Applications and Future Perspectives. *Drug Discov Today* **2016,** *21*, 288-298.

19. Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H. Deep-Learning-Based Drug–Target Interaction Prediction. *J. Proteome Res.* **2017,** *16*, 1401-1409.

20. Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; Ijzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a Chembl Bioactivity Benchmark Set. *J. Cheminform.* **2017,** *9*, 45.

21. Schneider, P.; Schneider, G. A Computational Method for Unveiling the Target Promiscuity of Pharmacologically Active Compounds. *Angew. Chem., Int. Ed. Engl.* **2017,** *56*, 11520-11524.

22. Irwin, J. J.; Gaskins, G.; Sterling, T.; Mysinger, M. M.; Keiser, M. J. Predicted Biological Activity of Purchasable Chemical Space. *J. Chem. Inf. Model.* **2018,** *58*, 148-164.

23. Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. Pass: Prediction of Activity Spectra for Biologically Active Substances. *Bioinformatics* **2000,** *16*, 747-748.

24. Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007,** *25*, 197-206.

25. Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X.-Q. Targethunter: An in Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *AAPS J.* **2013,** *15*, 395-406.

26. Gong, J.; Cai, C.; Liu, X.; Ku, X.; Jiang, H.; Gao, D.; Li, H. Chemmapper: A Versatile Web Server for Exploring Pharmacology and Chemical Structure Association Based on Molecular 3d Similarity Method. *Bioinformatics* **2013,** *29*, 1827-1829.

27. Liu, X.; Vogt, I.; Haque, T.; Campillos, M. Hitpick: A Web Server for Hit Identification and Target Prediction of Chemical Screenings. *Bioinformatics* **2013,** *29*, 1910-1912.

28. Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the Macromolecular Targets of De Novo-Designed Chemical Entities through Self-Organizing Map Consensus. *Proc. Natl. Acad. Sci.* **2014,** *111*, 4067-4072.

29. Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. Swisstargetprediction: A Web Server for Target Prediction of Bioactive Small Molecules. *Nucleic Acids Res.* **2014,** *42*, W32-W38.

30. Nickel, J.; Gohlke, B.-O.; Erehman, J.; Banerjee, P.; Rong, W. W.; Goede, A.; Dunkel, M.; Preissner, R. Superpred: Update on Drug Classification and Target Prediction. *Nucleic Acids Res.* **2014,** *42*, W26-W31.

31. Pogodin, P. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Pass Targets: Ligand-Based Multi-Target Computational System Based on a Public Data and Naive Bayes Approach. *SAR QSAR Environ. Res.* **2015,** *26*, 783-793.

32. Liu, X.; Gao, Y.; Peng, J.; Xu, Y.; Wang, Y.; Zhou, N.; Xing, J.; Luo, X.; Jiang, H.; Zheng, M. Tarpred: A Web Application for Predicting Therapeutic and Side Effect Targets of Chemical Compounds. *Bioinformatics* **2015,** *31*, 2049-2051.

33. Kringelum, J.; Kjaerulff, S. K.; Brunak, S.; Lund, O.; Oprea, T. I.; Taboureau, O. Chemprot-3.0: A Global Chemical Biology Diseases Mapping. *Database* **2016,** *2016*, bav123-bav123.

34. Awale, M.; Reymond, J. L. The Polypharmacology Browser: A Web-Based Multi-Fingerprint Target Prediction Tool Using Chembl Bioactivity Data. *J. Cheminform.* **2017,** *9*, 11.

35. Anusevicius, K.; Mickevicius, V.; Stasevych, M.; Zvarych, V.; Komarovska-Porokhnyavets, O.; Novikov, V.; Tarasova, O.; Gloriozova, T.; Poroikov, V. Synthesis and Chemoinformatics Analysis of N-Aryl-B-Alanine Derivatives. *Res. Chem. Intermed.* **2015,** *41*, 7517-7540.

36. Luo, M.; Wang, X. S.; Tropsha, A. Comparative Analysis of Qsar-Based Vs. Chemical Similarity Based Predictors of Gpcrs Binding Affinity. *Mol. Inf.* **2016,** *35*, 36-41.

37. Murtazalieva, K. A.; Druzhilovskiy, D. S.; Goel, R. K.; Sastry, G. N.; Poroikov, V. V. How Good Are Publicly Available Web Services That Predict Bioactivity Profiles for Drug Repurposing? *SAR QSAR Environ Res* **2017,** *28*, 843-862.

38. Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on Chembl. *Chem. Sci.* **2018,** *9*, 5441-5451.

39. Alberga, D.; Trisciuzzi, D.; Montaruli, M.; Leonetti, F.; Mangiatordi, G. F.; Nicolotti, O. A New Approach for Drug Target and Bioactivity Prediction: The Multi-Fingerprint Similarity Search Algorithm (Mussel). *J. Chem. Inf. Model.* **2018**, DOI: 10.1021/acs.jcim.8b00698.

40. Simonin, C.; Awale, M.; Brand, M.; van Deursen, R.; Schwartz, J.; Fine, M.; Kovacs, G.; Häfliger, P.; Gyimesi, G.; Sithampari, A.; Charles, R. P.; Hediger, M.; Reymond, J. L. Optimization of Trpv6 Calcium Channel Inhibitors Using a New 3d Ligand Based Virtual Screening Method. *Angew. Chem., Int. Ed. Engl.* **2015,** *54*, 14748-14752.

41. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010,** *50*, 742-754.

42. Raevsky, O.; Sapegin, A.; Zefirov, N. The Qsar Discriminant-Regression Model. *Quant. Struct.-Act. Relat.* **1994,** *13*, 412-418.

43. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The Chembl Bioactivity Database: An Update. *Nucleic Acids Res.* **2014,** *42*, D1083-D1090.

44. Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **2009,** *4*, 1803-1805.

45. van Deursen, R.; Blum, L. C.; Reymond, J. L. A Searchable Map of Pubchem. *J. Chem. Inf. Model.* **2010,** *50*, 1924-1934.

46. Blum, L. C.; van Deursen, R.; Reymond, J. L. Visualisation and Subsets of the Chemical Universe Database Gdb-13 for Virtual Screening. *J. Comput.-Aided Mol. Des.* **2011,** *25*, 637-647.

47. Ruddigkeit, L.; Blum, L. C.; Reymond, J. L. Visualization and Virtual Screening of the Chemical Universe Database Gdb-17. *J. Chem. Inf. Model.* **2013,** *53*, 56-65.

48. Probst, D.; Reymond, J. L. Fun: A Framework for Interactive Visualizations of Large, High Dimensional Datasets on the Web. *Bioinformatics* **2017,** *37*, 1433-1435.

49. Awale, M.; Reymond, J. L. Atom Pair 2d-Fingerprints Perceive 3d-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of Zinc and Gdb-17. *J. Chem. Inf. Model.* **2014,** *54*, 1892-1897.

50. Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. *Methods Mol. Biol.* **2011,** *672*, 39-100.

51. Probst, D.; Reymond, J. L. Smilesdrawer: Parsing and Drawing Smiles-Encoded Molecular Structures Using Client-Side Javascript. *J. Chem. Inf. Model.* **2018,** *58*, 1-7.

Graphics for the Table of Contents: